Price and Service Discrimination in Queuing Systems: Incentive Compatibility of G$c\mu$ Scheduling

Jan A. Van Mieghem
Kellogg Graduate School of Management, Northwestern University, Evanston, Illinois 60208
vanmieghem@kellogg.northwestern.edu

This article studies the optimal prices and service quality grades that a queuing system—the "firm"—provides to heterogeneous, utility-maximizing customers who measure quality by their experienced delay distributions. Results are threefold: First, delay cost *curves* are introduced that allow for a flexible description of a customer's quality sensitivity. Second, a comprehensive *executable approach* is proposed that analytically specifies scheduling, delay distributions and prices for arbitrary delay sensitivity curves. The tractability of this approach derives from porting heavy-traffic Brownian results into the economic analysis. The generalized $c\mu$ (G$c\mu$) scheduling rule that emerges is dynamic so that, in general, service grades need not correspond to a static priority ranking. A benchmarking example investigates the value of differentiated service. Third, the notions of *grade* and *rate* incentive compatibility (IC) are introduced to study this system under asymmetric information and are established for G$c\mu$ scheduling when service times are homogeneous and customers atomistic. Grade IC induces correct grade choice resulting in perfect service discrimination; rate IC additionally induces centralized-optimal rates. Dynamic G$c\mu$ scheduling exhibits negative feedback that, together with time-dependent pricing, can also yield rate incentive compatibility with heterogeneous service times. Finally, *multiplan pricing*, which offers all customers a *menu* with a choice of multiple rate plans, is analyzed.

(*Pricing*; *Quality of Service* (*QoS*); *Differentiation*; *Queuing*; *Incentive Compatibility*; *Asymmetric Information*; *Delay Costs*; *Scheduling*; *Dynamic Priority*; *Generalized $c\mu$ Rule*; *Threshold Rules*)

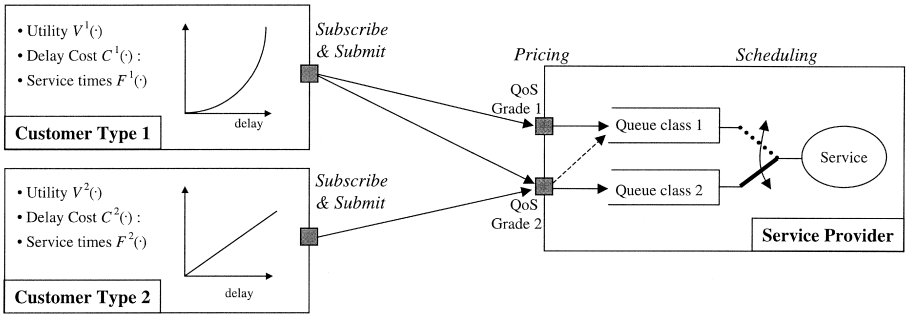
1. Introduction and Summary of Results

A service provider is considering offering differentiated quality of service to a market consisting of several customer segments or *types*. Quality of service (QoS) is measured by the delay distributions that customers experience in receiving service. Differentiation derives from offering multiple *service grades* that each render a different delay distribution at a different price. The provider has three direct controls that together define the *mechanism* used to achieve differentiation: the number of grades, their price schedules, and a scheduling rule, which determines the order in which service requests are served. We will analyze how this mechanism can be designed to tailor the decentralized allocation of scarce processing resources to individually-acting customers.

The example shown in Figure 1 illustrates our model as follows. Customers can sign up as "subscribers" to receive access rights to send a stream of service

0025-1909/00/4609/1249$05.00
1526-5501 electronic ISSN

Figure 1 **A Differentiated Quality of Service Model Where Customer Types Choose Service Grades and Submission Rates, and the Service Provider Decides on the Number of Service Grades, Their Prices, and Scheduling**



requests or "jobs" to one or multiple service grades over time. A customer type is multidimensional and characterized by a triplet of functions: a gross utility function specifies the value that a customer derives from service, a delay cost function models QoS sensitivity by specifying the value degradation that accompanies longer delay, and a service time distribution function. Customers choose service grades and submission rates that maximize their net utility, which is gross utility minus delay costs and price. The decision structure considers two objectives for the service provider: maximize total system utility (social pricing) or its own server profits (monopoly pricing).

This article strives to contribute along three dimensions. The first dimension concerns *model formulation*: this model introduces delay cost *curves* that allow a flexible description of a customer's quality sensitivity, as assumed for type 1 in Figure 1. Virtually all of the literature has restricted attention to linear delay costs—i.e., one marginal cost *number* per type—and associated static priority scheduling rules. Including delay cost curves is an important addition because delay sensitivity is nonlinear in many practical settings, such as telephone and Internet service or where lead-times or due-dates are concerned. The model also presents an integrated approach to service design, pricing, and execution. Conventional models take QoS levels as exogeneously given, usually in the form of hard QoS guarantees quoted to customers as in Maglaras and Van Mieghem (2000). Here, delay cost functions are the building block to value QoS, and differentiation and associated delay distributions emerge endogeneously from profit-maximizing firm behavior. Associated with each service grade, then, is a price and delay distribution measuring the expected, but not guaranteed, QoS level.

The second dimension concerns *mode of analysis* and *solution technique*. Optimal service levels are determined by an optimal scheduling rule, which is unknown for a general delay cost structure. Instead of restricting attention to exact optimality and linear delay costs, this article proposes to approximate the optimal unknown policy by a concrete scheduling rule, called the generalized $c\mu$ ($Gc\mu$) rule, which is asymptotically optimal in heavy traffic as shown in Van Mieghem (1995). Moreover, $Gc\mu$ is a dynamic scheduling rule so that, in general, service grades need not correspond to a static priority ranking. By importing simple heavy-traffic Brownian results, this approximate mode of analysis makes complex stochastic, economic systems tractable. Its power is manifested by its results: We present an executable proposal that specifies dynamic scheduling policies and its associated delay distributions and prices analytically for arbitrary delay sensitivity curves. As an example, we analyze the value of offering differentiated service. The model is also sufficiently general to serve as a first step in extending the analysis to a network setting, for which our approximate mode of analysis should prove useful.

Finally, the third dimension concerns the study of *price and service discrimination under full and asymmetric information*. We analyze three cases, listed in decreasing order of information availability and allowed actions (and thus decreasing performance):

1. In the *centralized system* the service provider has full information and can directly control all customer

rates. Optimal scheduling differentiates by customer type resulting in perfect service discrimination. In queuing terminology this means that *queuing classes*, which contain the finest information set on which a scheduling rule can be defined, correspond to customer types. The centralized system yields an upper bound on the performance of decentralized systems in which the service provider uses grades and prices to indirectly control customer-chosen grades and rates.

2. Under *full information*, the server observes each job's originating customer type and can implement perfect service discrimination, regardless as to which grades customers choose.[1] With correct allocation of job types to queuing classes, as indicated by the dotted line in Figure 1, prices only need to influence the total rate into each class. If the server can set customer-specific prices, offering and pricing one grade is sufficient to induce each type to choose the centralized-optimal rate. Such a mechanism in which customers self-select the centralized-optimal rate is called *rate incentive-compatible* (*IC*). It perfectly coordinates the decentralized system and achieves perfect price discrimination. If the server cannot set customer-specific prices, it can still replicate the former outcome by offering as many grades as there are types and have grades correspond to queuing classes. Under full information, scheduling control by itself[2] can induce customers to each self-select their centralized-optimal grade, which is called *grade incentive compatibility*. Grade-specific prices can then again be set only to induce the centralized optimal rates. Section 3 shows that with full information, customer reservation prices can be extracted by a customer-specific two-part tariff, consisting of a fixed subscription fee and variable usage fee.

3. Under *asymmetric information*, the service provider cannot observe the type of a job. Grade information now is the finest information on which the server can price and schedule, which typically leads to *imperfect* service and price discrimination. (For example, in Figure 1, queuing class two may hold customers of either type and the server cannot distinguish between them.) Under asymmetric information, pricing and scheduling interact to induce grade and rate choices and grade incentive compatibility can no longer be achieved through scheduling only. Thus, grade and rate IC are much harder to achieve than under full information where the two can be effectively separated. In special cases, however, the mechanism may achieve the dual goal of inducing customers not only to choose the "right" grade, but also the "right" rate to that grade. We establish such grade and rate incentive compatibility for G$c\mu$ scheduling when customers are atomistic and have homogeneous service time distributions. This directly extends the results of Lederer and Li (1997) and Mendelson and Whang (1990) for static priority queuing to dynamic G$c\mu$ scheduling with arbitrary delay cost functions. We explain how negative feedback inherent in dynamic G$c\mu$ scheduling and time-dependent pricing can reinforce each other to also yield rate incentive compatibility when service time distributions are type-dependent. Finally, instead of offering all customers a single price plan the firm can offer a *menu* with a choice of multiple rate plans, known as *multiplan pricing* and widely adopted in practice. The analysis of multiplan pricing in economic queuing models with asymmetric information appears to be novel.

The outline of this article is as follows. This introduction concludes with a short literature review below. Section 2 presents the model. Section 3 analyses Cases 1 and 2: the centralized system followed by full information. Section 4 shows how to derive the G$c\mu$ scheduling rule and its delay distributions. Section 5 illustrates this approach by valuing differentiated service using the G$c\mu$ rule versus traditional fixed priority and FIFO service. Section 6 analyzes the third case of asymmetric information. Section 7 offers concluding remarks.

Lederer and Li (1997) and Mendelson and Whang (1990) provided our main starting inspiration. The literature that studied the use of pricing to manage the impact of externalities in congestion systems is extensive and appears to have been started by Naor

[1] Discriminatory service highlights scheduling as a valuable lever to improve price discrimination in backoffice operations such as e-commerce or call-centers where the server may have full type information.

[2] For example, willfully delaying any type i job submitted to grade $j \neq i$ for a long time inflicts such high delay cost onto type i that would discourage such choice of grade.

(1969). Knudsen (1972) and Lippman and Stidham (1977) first highlighted the difference between profit-maximizing and socially optimal control of queuing systems. In a hallmark paper, Mendelson (1985) embedded the queuing system in an economic framework. Dolan (1978), Mendelson and Whang (1990), and Rao and Petersen (1998), among others, study incentive-compatible pricing of static priority queues. De Vany and Saving (1983), Reitman (1991), Loch (1991), Lederer and Li (1997), and Cachon and Harker (1999) consider delay-quality differentiation in competitive industry models. Lui (1985) and Ha (1998) add service-rate effort as an additional decision variable to incentive-compatible pricing for homogeneous customers under FIFO scheduling. Bradford (1996), adds static routing control. Afèche and Mendelson (2000) take a first crack at the network extension. Ha (1999) shows that a single variable price is optimal and incentive compatible for heterogeneous customers that choose service requirements under uniform processor sharing. Finally, we refer to Courcoubetis (1998) and Gibbens and Kelly (1999) for an overview of the Internet-related pricing literature, which seems fairly disconnected from the above.

2. A Multitype, Multiservice Subscription Model

A basic element of the formulation is a model of the heterogeneity of customer behavior and of service offerings. For this purpose, each customer or market segment is classified as one of several *types*, indicated by superscripts $i=1,\ldots, m$, while the different service grades are indicated by subscripts $k=1, \ldots, n$.

Customer Behavior Modeling. Customers can sign up as subscribers to receive access rights to send a stream of service requests or "jobs" to the service provider over time. They can distribute their total service needs over the various service grades. We model customer type i's stream of service requests as a renewal vector process with average rate of requests sent to grade k denoted by λ_k^i. The n-dimensional subscription rate vector $\lambda^i=(\lambda_1^i, \ldots, \lambda_k^i, \ldots, \lambda_n^i)$ represents the strategic decision variable of customer type i.

Type i customers are identified by a triplet of functions (V^i, C^i, F^i) defined as follows. Receiving services at total rate $\lambda_+^i = \sum_k \lambda_k^i$ generates gross utility or *value* $V^i(\lambda_+^i)$ per unit of time to customer type i. At the same time, any delay in receiving service may degrade that value. That is, a particular type i job may have to wait some time t before service is initiated, inflicting a *delay cost* $C^i(t)$ onto customer type i. The actual processing time needed to serve a request is called the *service time*. As usual, service times are assumed to be iid random variables with generic representative τ^i with cumulative distribution function F^i.

Customers may split their total service needs over the various service grades, precisely because a "better" grade may provide more timely service and reduce delay costs. Customer i's delay experienced by a service request to grade k is modeled by the generic random variable t_k^i, because delays may depend on quantities that are unknown at the time when customer i makes her subscription decision. Thus, customer i will make her subscription rate decision λ_k^i to grade k anticipating a total grade k delay cost rate $\lambda_k^i \mathrm{E}C^i(t_k^i)$, where E denotes the expectation operator under equilibrium conditions (a precise notion of equilibrium will be discussed later). After processing a job, a customer must pay a price that may be a function of the *actual* (ex-post observed) processing time τ^i and of the chosen grade. In addition, depending on the *regulatory environment*, the price may be customer-specific. If customer-specific pricing is not allowed, then all customers must be offered the same price contract. In general, then, customer type i's ex-ante expected payment *rate* (per unit of time) can be denoted by $\mathrm{E}P^i(\lambda^i, \tau^i)$, which we simplify to $P^i(\lambda^i)$ if the charge is not explicitly service-time dependent and to $\mathrm{E}P(\lambda^i, \tau^i)$ if customer-specific pricing is not allowed. In equilibrium, customer i will make her subscription rate decision[3] λ^i to maximize

[3] We first assume that a single decision maker sets the subscription vector λ^i for customer type i. This implies that each customer type either represents one major customer (e.g., corporate customer of an IT service provider) or many customers who *collusively* set λ^i. Later, in §6, we simplify to the "atomistic model" where a type comprises many small customers who each decide individually whether or not to subscribe at an infinitesimal rate and the type vector λ^i is the aggregate result of these individual decisions.

her expected net monetary profit *rate* π^i, where

$$\pi^i = V^i(\lambda_+^i) - \sum_{k=1}^{n} \lambda_k^i \mathrm{E}C^i(t_k^i) - \mathrm{E}P^i(\lambda^i, \tau^i), \tag{1}$$

which is the value rate net of expected delay costs and payments. All value functions are concave increasing, the traditional economic assumption of decreasing marginal returns. The delay cost functions are assumed to be convex increasing, reflecting the fact that more waiting is increasingly costly.

The Process View. In addition to choosing the number n of service grades, the service provider designs its service offering through two strategic control levers: pricing and scheduling. The service provider can partially control subscription rates by setting the *price* schedules $\{P^i(\lambda, t) : i = 1, \ldots, m\}$ defined earlier. Indeed, this more tailored control is exactly the motive behind offering differentiated services. In addition, the service provider has *dynamic internal* control in that it can choose a scheduling rule r to decide at each point in time how to serve jobs through its service process. (The firm has no explicit admission control because once customers have subscribed, their jobs must be served.) The service process is modeled by a *queuing process* with q classes, indexed $j = 1, \ldots, q$. The class designation captures the finest possible information that the service provider possesses: He cannot distinguish ex-ante among different jobs in class j and, hence, must treat them homogeneously. Consequently, the scheduling rule r is defined in terms of classes, which may be thought of as physical queues. Under full information, the service provider can observe each job's type and classes are grade and customer specific: class $j = (i, k)$ and its arrival process has rate $\Lambda_j = \lambda_k^i$ and its service times are i.i.d. with distribution $F_j = F^i$ and mean denoted $m_j = 1/\mu^i$. Customer i's generic random delay time t_k^i for grade k equals class j delay time t_k. Under asymmetric information, types are not observable and classes are grade-specific only: Class $j = k$ aggregates all subscriptions to service grade k. Its arrival process thus is a compound renewal process with rate $\Lambda_j = \sum_i \lambda_j^i$ and its service times distribution becomes

$$F_j(x) = \sum_{i=1}^{m} \frac{\lambda_j^i}{\Lambda_j} F^i(x) \quad \text{with mean} \quad m_j = \frac{1}{\mu_j} = \sum_{i=1}^{m} \frac{\lambda_j^i}{\Lambda_j} \frac{1}{\mu^i}. \tag{2}$$

In that case, all customers submitting to grade k receive the same generic random delay time t_k. Under either information structure, the average rate of work (or "traffic intensity") submitted to class j and the total system is, respectively,

$$\rho_j = \frac{\Lambda_j}{\mu_j}, \quad \text{and} \quad \rho = \sum_{j=1}^{q} \rho_j. \tag{3}$$

We consider dynamic scheduling rules that may depend on the current internal state of the queuing process and the arrival vector Λ because the service provider (but not the customers) observes the dynamic queue-count in each class. We do restrict attention, however, to stationary control rules that do not explicitly depend on time. Then, regardless of the queuing system's internal complexity, the service process can be summarized by a *technology function*, which specifies how the control rule r transforms a total subscription arrival vector Λ into class-dependent delay times, represented by their distribution functions $F_j^r(\cdot \mid \Lambda)$. Thus, customer i's expected delay cost when submitting a job to grade k is:

$$\mathrm{E}_\Lambda^r C^i(t_k^i) = \begin{cases} \int C^i(t) dF^r_{j=(i,\,k)}(t|\Lambda) \\ \quad \text{under full information,} \\ \int C^i(t) dF^r_{j=k}(t|\Lambda) \\ \quad \text{under asymmetric information,} \end{cases} \tag{4}$$

where E_Λ^r represents the expectation operator when rule r is used and the total arrival vector is Λ. Two remarks are at place here: First, (4) shows that customer i's delay cost of grade k depends on the total subscription vector Λ. Reversing the argument, by sending a job to grade k, customer i may impact the delay cost of other customers. This *externality* effect will impact the pricing decisions. Second, (4) allows for *perfect service discrimination* in the sense that scheduling is customer-type-specific if types are observable so that different customers submitting to the same grade can receive a different expected QoS.

The service provider bears an operating cost $C^S(\Lambda)$ per unit of time when processing vector Λ. As traditionally, we assume that C^S is convex

increasing. In equilibrium, using price rate schedules $\{P^i(\lambda, t): i=1, \ldots, m\}$ and internal control rule r, the service provider will earn profit rate π^S, where

$$\pi^S = \sum_{i=1}^{m} \mathrm{E}P^i(\lambda^i, \tau^i) - C^S(\Lambda). \tag{5}$$

We distinguish two objectives for the service provider: maximize either its own individual profit π^S or social systemwide profits $\pi^S + \sum_i \pi^i$.

We assume the following *information structure*: aggregate type characteristics $\{(V^i, C^i, F^i): i=1, \ldots, m\}$ and the mechanism defined by the triplet (n, P, r) of number of grades, their price schedules, and a stationary scheduling rule (or, equivalently, the delay distributions F_k^r), are common information to all agents (customers and service provider). While each customer always knows its type, we distinguish between *full* and *asymmetric information*, depending on whether the service provider can observe a job's originating customer type or not. As said earlier, the queue-count state vector of the internal service process is observable to the service provider but not to the customers; grades and prices form the only interface between service system and customers. As usual, actual service times τ^i are not observable ex-ante.

This information structure guarantees that the noncooperative game is well specified.[4] We will analyze the system dynamics in equilibrium, which we define in the usual Nash sense as follows. An *equilibrium* of this noncooperative decision model is any quadruplet (n, P, r, λ) of number of grades, price schedules, scheduling rule, and customer subscription rate vectors that satisfies the system dynamics and that is such that no agent has an incentive to unilaterally deviate from the equilibrium. Such equilibrium therefore captures a consistent and sustaining solution to the service provider's and the m customers' decision problems.

[4] Common information requirements may be reduced at the additional complication of modeling learning and dynamic pricing. See Masuda and Whang (1999) for a first exploration of such approach.

3. Price and Service Discrimination Under Full Information

Centralized System. It is instructive to first analyze the relaxed problem where one central planner makes all decisions (which implies full information) to maximize social, systemwide profits $\Pi^r(\lambda)$, where

$$\Pi^r(\lambda) = \pi^S + \sum_i \pi^i = \sum_{i=1}^{m} (V^i(\lambda_+^i) - \sum_{k=1}^{n} \lambda_k^i \mathrm{E}_\Lambda^r C^i(t_k^i)) - C^S(\Lambda). \tag{6}$$

The performance of this centralized system represents a "first-best" solution and provides an upper bound to the decentralized system performance. The central planner has direct control of the scheduling rule r and the rate vectors λ^i, obviating indirect control through pricing *and* offering grades. Indeed, "service grades," in the sense defined earlier, are superfluous in the centralized system: The central planner allocates types directly to queuing classes as that offers the finest information set on which can be scheduled.[5] (Such allocation clearly dominates any "mixing" of several customer type flows into one class, as the finer allocation can always replicate the coarser.) Let t_j denote the generic random delay of class j. Let $v^i = (\partial/\partial\lambda_k^i)V^i$ represent customer i's marginal value rate function and $c_k = (\partial/\partial\Lambda_k)C^S = (\partial/\partial\lambda_k^i)C^S$ the service provider's marginal cost of class k subscriptions. It is obvious that:

PROPOSITION 1. *The optimal centralized system yields profits* $\Pi^* = \Pi^{r^*}(\lambda^*)$ *by using a scheduling rule* r^* *that minimizes total delay cost rate, denoted by* $DC_\Lambda^r = \sum_{i,k} \lambda_k^i \mathrm{E}_\Lambda^r C^i(t_k^i)$ *for any given load vector* Λ, *and achieves perfect service discrimination* (*queuing classes are type-specific*: $\lambda_{k\neq i}^i = 0$). *Type* i*'s optimal scalar rate* $\lambda_i^* \geq 0$ *and*

[5] If the service provider is ex-ante restricted to $q < m$ queue-classes, then at least one class will contain multiple customer types and scheduling will be coarser (imperfect service discrimination). In that case, optimal allocation of customer types to classes solves (7) after replacing the single summation by $\sum_{j=1}^{m} \sum_{l=1}^{q} \lambda_l^{*j} \partial/\partial\lambda_k^i \mathrm{E}_{\Lambda^*}^{r^*} C^j(t_l) + u_k^{*i}$, where $u_k^{*i} \leq 0$ is the optimal Lagrange multiplier on the nonnegativity constraint of $\lambda_k^i : \lambda_k^{*i} u_k^{*i} = 0$.

dual variable $u_i^* \le 0$ *solve*: $\forall i=1, \ldots, m: \lambda_i^* u_i^* = 0$ *and*[6]

$$v^i(\lambda_i^*) = \mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i) + \sum_{j=1}^{m} \lambda_j^* \frac{\partial}{\partial \lambda_i} \mathrm{E}_{\Lambda^*}^{r^*} C^j(t_j) + c_i(\Lambda^*) - u_i^*. \tag{7}$$

Convexity of the total expected delay cost $\sum_i \lambda_i \mathrm{E}_{\Lambda}^{r^*} C^i(t_i)$ would yield concavity of Π^{r^*} and thus uniqueness of the solution. This is the case with linear delay cost functions $C^i(t) = c^i t$ for which the optimal scheduling rule r^* is the $c\mu$ rule, which ranks the priority of classes in the order of index $c^i \mu_i$. For general delay cost functions, however, the optimal scheduling rule and its delay distributions are unknown. Hence, convexity and uniqueness, while plausible because delays are typically convex in the rates and the delay cost functions are convex, cannot be guaranteed.

The necessary first-order conditions (7) have a familiar economic interpretation: the optimal rate λ_i^* equates the marginal value of customer i's incremental class i job with its total marginal cost. The latter is born by three parties: (1) the "self-regulating" term $\mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i) + \lambda_i^* (\partial/\partial \lambda_i) \mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i)$ is born by customer i, (2) the externality term $\sum_{j \neq i} \lambda_j^* (\partial/\partial \lambda_i) \mathrm{E}_{\Lambda^*}^{r^*} C^j(t_j)$ is inflicted onto other customers, and (3) the marginal operating cost $c_k(\Lambda^*)$ is born by the service provider. Finally, a zero rate $\lambda_i^* = 0$ obtains if its marginal value does not outweigh its marginal cost.

Full Information and Customer-Specific Pricing. When the service provider observes types, grades are superfluous for scheduling purposes and perfect service discrimination is achieved by allocating customer types to classes one-to-one. Yet, the service provider needs at least one grade because service grades and their prices form the observable interface with customers in the decentralized system. With optimal scheduling already guaranteed, prices only need to induce customers to self-select the centralized-optimal rate, which is called *rate incentive compatibility* (*IC*).

[6] The notation of λ_i^i and $\partial/\partial \lambda_i^i$ will be simplified to λ_i and $\partial/\partial \lambda_i$ whenever possible.

Proposition 2. *Let* (λ^*, r^*) *represent the centralized optimum and define* $\forall i = 1, \ldots, m$:

$$b_i^* = c_i(\Lambda^*) + \sum_{j=1,\ j \neq i}^{m} \lambda_j^* \frac{\partial}{\partial \lambda_i} \mathrm{E}_{\Lambda^*}^{r^*} C^j(t_j), \tag{8}$$

$$a_i^* = V^i(\lambda_i^*) - \lambda_i^* \mathrm{E} C^i(t_i) - b_i^* \lambda_i^*. \tag{9}$$

Under full information and customer-specific pricing, it is sufficient to offer one service grade: $n^* = 1$. *If the delay cost* $\lambda_i \mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i)$ *is locally* (*globally*) *convex in the scalar* λ_i *at* λ^* *for each* i, *then an equilibrium exists* (*is unique*) *and any customer-specific affine pricing* $P^{i*}(\lambda) = a_i + b_i^* \lambda$, *where* $a_i \le a_i^*$, *together with* (n^*, λ^*, r^*), *is a social-welfare equilibrium. If* $a_i = a_i^*$, *it is also a profit-maximizing equilibrium and the service provider extracts all system profits*: $\pi^{*S} = \Pi^*$ *and* $\pi^{*i} = 0\ \forall i$.

Remarks. The prices P^{i*} are *two-part tariffs* where a_i is the fixed subscription fee that gives customer i access rights, while b_i^* is the variable price for a unit rate from customer i. These constant marginal prices force customer i to incorporate the marginal externality that she inflicts onto other agents in her decision making process. Section 6 will show that the fixed fee is zero in the simpler atomistic model and that more complex schedules, which may be service-time dependent, will be used under asymmetric information. The prices P^* are *perfectly discriminating* because each customer is charged its "reservation price": the service provider extracts all surplus and customers are indifferent between participating or not. Finally, this mechanism perfectly *coordinates* the decentralized system because it induces the centralized-optimal solution.

Proof. Let us briefly review the standard economic argument to derive the equilibrium $(n^*, P^*, \lambda^*, r^*)$ under full information as it will be useful for §6. If an equilibrium exists, customer i will make her subscription decision λ^i to maximize her expected net profit rate π^i as given in (1) with necessary first-order conditions:

$$v^i(\lambda_+^i) = \mathrm{E}_{\Lambda}^{r} C^i(t_k) + \sum_{j=1}^{n} \lambda_j^i \frac{\partial}{\partial \lambda_k^i} \mathrm{E}_{\Lambda}^{r} C^i(t_j^i) + \frac{\partial}{\partial \lambda_k^i} P^i(\lambda^i) + u_k^i \qquad \forall k = 1, \ldots, n, \tag{10}$$

and $\lambda_k^i u_k^i = 0$, where $u_k^i \le 0$ is the Lagrange multiplier on the nonnegativity constraint of λ_k^i. Compare the

decentralized conditions (10) with the centralized conditions (7). To induce rate IC, it suffices to offer one grade: $n^*=1$ so that λ^i becomes a scalar. In addition, the service provider must use the centralized-optimal control rule r^* and a price schedule with constant marginal prices $b_i^*=(d/d\lambda^i)P^i(\lambda^i)$ to induce the rate λ_i^*. Then, under the local convexity of $\lambda^i \mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i)$, customer i has no incentive to unilaterally deviate from λ_i^*. Second, the participation conditions insure that customer i will "participate": $a_i \le a_i^*$ implies $\pi^i(\lambda^{*i}) \ge 0$ so that customer i is at least indifferent to subscribing or not.

Similarly, we must show that, if all customers choose the centralized solution λ^*, then the service provider will choose one grade, pricing schedule P^* and scheduling rule r^*, and that he has no incentive to deviate from that decision. Clearly, the sum of all profit rates in the decentralized system cannot exceed the optimal centralized system profit: $\sum_i \pi^i + \pi^S \le \Pi^*$. Thus, the best the service provider can possibly hope for is to capture all the profits of the system: $\pi^{*S}=\Pi^*$ and $\sum_i \pi^i=0$, which implies $\pi^i=0$ $(\forall i)$ to satisfy our participation constraint. This indeed is possible if the service provider uses one grade with control rule r^* and pricing schedule P^* with intercept a_i^*. The service provider's profit rate becomes:

$$\begin{aligned}\pi^S &= \sum_i P^{*i}(\lambda^{*i}) - C^S(\Lambda^*) \\ &= \sum_i [V^i(\lambda^{*i}) - \lambda_i^* \mathrm{E}C^i(t_i)] - C^S(\Lambda^*) = \Pi^*,\end{aligned}$$

so that service provider has no incentive to unilaterally deviate from his decision. Under social welfare maximization, the allocation of system profits Π^* to agents is irrelevant so that any $a_i \le a_i^*$ forms an equilibrium. Finally, if all customer delay costs $\lambda_i \mathrm{E}_{\Lambda^*}^{r^*} C^i(t_i)$ are globally convex, all first-order conditions have a unique solution λ^*. □

Full Information and Grade-Specific Pricing. If the service provider cannot set customer-specific prices, he can still induce the centralized outcome by offering as many grades as there are types ($n^*=m$) and by making queuing classes grade specific. Under full information, it is easy to induce customers to each self-select their centralized-optimal grade (class) and achieve such *grade incentive compatibility*. The *grade IC conditions* can be stated as $b_{j\ne i}^i(\lambda) < b_i^i(\lambda)$, where

$$b_j^i(\lambda) = v^i(\lambda_i^i) - \mathrm{E}_\Lambda^{r^*} C^i(t_j^i) - \lambda_i \frac{\partial}{\partial \lambda_j^i} \mathrm{E}_\Lambda^{r^*} C^i(t_i^i) - b_j \qquad (11)$$

is type i's marginal profit at λ from a grade j subscription; the *rate IC conditions* require in addition that $b_i^i(\lambda^*)=0$. Under full information, grade IC can be achieved purely through scheduling. For example, willfully delaying a type i job submitted to grade $j \ne i$ for a constant long time T^i, where $C^i(T^i) > v^i(\lambda_i^{*i})$, inflicts a high delay cost onto type i that, exacerbated by the fixed fee, would discourage such choice of grade: $b_j^i(\lambda^*) < 0$, which makes $\lambda^{*i}=(0, \ldots, 0, \lambda_i^*, 0, \ldots, 0)$ an equilibrium. With grade IC, grade-specific prices become equivalent to customer-specific prices. As before, it is optimal to set them at b_i^* to induce the centralized optimal rates. In conclusion: under full information, differentiated service grades and scheduling achieve grade IC, leaving prices the only task of inducing rate IC, which results in perfect price discrimination and coordination.

4. The Gcμ Scheduling Rule and its Delay Distributions

The previous section highlighted the role of the optimal scheduling rule r^*: it minimizes total expected delay costs, denoted by $DC_\Lambda^{r^*} = \sum_{i,k} \lambda_k^i \mathrm{E}_\Lambda^r C^i(t_k^i)$, for any given load vector Λ. In addition, we need its delay distributions to quantify the expected delay cost and its gradients to compute the optimal rates λ^* and prices P^*. Unfortunately, as described in the introduction, the exactly-optimal scheduling rule is unknown when delay cost functions are convex increasing. In Van Mieghem (1995), however, we present the Generalized $c\mu$ (G$c\mu$) rule, which is shown to be asymptotically optimal in a single server system in "heavy traffic" (i.e., if traffic intensity $\rho \to 1$). Here we show how to specify the G$c\mu$ rule and its delay distributions in our setting and we provide some intuition behind the rule. To stress the approximate mode of analysis, we will use $\simeq$ to denote approximate relationships that are asymptotically exact if $\rho \to 1$.

DEFINITION. G$c\mu$ scheduling serves class j jobs in the order they arrive (FIFO). Classes correspond to customer types if types are observable, otherwise

they correspond to grades. Let $c^i(x) = (d/dt)C^i(x)$ be customer type i's marginal delay cost and $N_j(t)$ the number of waiting class j jobs at time t. Then Gcμ serves the class with highest priority index

$$I_j(t) = \begin{cases} c^j\left(\dfrac{N_j(t)}{\Lambda_j}\right)\mu_j & \text{under full information,} \\ & \forall j = 1, \ldots, m\colon\ \Lambda_j = \displaystyle\sum_{k=1}^{n} \lambda_k^j, \\ \displaystyle\sum_{i=1}^{m} \frac{\lambda_j^i}{\rho_j} c^i\left(\dfrac{N_j(t)}{\Lambda_j}\right) & \text{under asymmetric} \\ & \text{information,} \\ & \forall j = 1, \ldots, n : \Lambda_j = \displaystyle\sum_{i=1}^{m} \lambda_j^i. \end{cases} \tag{12}$$

Notice that the Gcμ rule is well defined, even under asymmetric information, as will be illustrated in the next section. (If types are not observable, the server can infer the rates λ_j^i under rational decision making by the customers just like we did for Proposition 2). To specify delay distributions, we need a mapping $g(\cdot)$ from $\mathbb{R}_+$ to "class-space" $\mathbb{R}_+^q$ that is defined as the solution of the following linearly constrained convex optimization problem for $W \geq 0$:

$$g(W) = \arg \min_{w \in \mathbf{R}_+^n} \sum_{i=1}^{m} \sum_{k=1}^{n} \lambda_k^i C^i(t_k^i) \tag{13}$$

$$\text{s.t.:} \begin{cases} \displaystyle\sum_{j=1}^{m} w_j = W \text{ and } \forall i, k\colon\ t_k^i = \frac{w_i}{\rho_i}, \\ \quad \rho_i = \displaystyle\sum_{k=1}^{n} \frac{\lambda_k^i}{\mu^i} \text{ under full information,} \\ \displaystyle\sum_{j=1}^{n} w_j = W \text{ and } \forall i, k\colon\ t_k^i = \frac{w_k}{\rho_k} \\ \quad \rho_k = \displaystyle\sum_{i=1}^{m} \frac{\lambda_k^i}{\mu^i} \text{ under asymmetric information.} \end{cases} \tag{14}$$

As the next section will illustrate, this mapping defines the *switching curve* of the Gcμ rule parameterized by scalar W. Denote the aggregate squared coefficient of variation of the service and interarrival times by C_s^2 and C_a^2, respectively.

PROPOSITION 3. *The Gcμ rule is asymptotically optimal for all arrival vectors* Λ *such that* $\rho(\Lambda) \to 1$. *Its delay distributions are* $F_j^{Gc\mu}(t|\Lambda) \simeq F_W(g_j^{-1}(\rho_j t)|\Lambda)$, *where the mapping* g *solves* (13) *and* (14) *and*

$$F_W(t|\Lambda) \simeq 1 - \rho e^{-\gamma t}, \quad \textit{where} \quad \gamma^{-1} = \mu^{-1} \frac{\rho}{1-\rho} \frac{C_a^2 + C_s^2}{2}. \tag{16}$$

(The Appendix gives the intuition behind Proposition 3, whereas Van Mieghem (1995) contains precise formulations and proofs.) The mixed distribution (16) is one of many possible Brownian approximations to the total workload distribution of a GI/G/1 queue. We use it, rather than the asymptotically equivalent $1 - e^{-\gamma t}$, because it makes our approach consistent with the exact results for the single-class FIFO M/M/1 queue, for which $g(W) = W$ and

$$F^{FIFO}(t) = 1 - \rho e^{-\mu(1-\rho)t} = F_{\text{single\_class}}^{Gc\mu}(t).$$

While the Gcμ rule minimizes total delay costs asymptotically in heavy traffic, our approximate mode of analysis derives from the fact that in this model arrival rates and traffic intensity ρ are endogeneous. Cost minimization will ensure moderate traffic in equilibrium and we propose to use the Gcμ rule in this regime for which optimality has not been proved in general. The benchmarking study in the next section and our simulation analysis in Van Mieghem (2000a), however, show that the approximation is very good for moderate traffic. Moreover, the rule is *prima facie* reasonable and appealing for implementation and for complex problems such as service time tail objectives (Ayhan and Olsen 2000) or lead-time constraints (Van Mieghem 2000b).

5. Pricing and Valuation of Differentiated Service: An Example

This section illustrates our methodology with a comprehensive benchmarking example assuming types are observable; the next section considers the case of asymmetric information. The executable proposal

is: (i) Characterize the Gcμ rule for the delay cost functions at hand. (ii) Calculate the associated delay distributions and total expected delay cost DC^*. (iii) Solve equations (7) for the centralized optimal rates λ^*. (iv) Evaluate the gradients of DC^* at λ^* to derive the equilibrium prices P^* as in Proposition 2. Verify local convexity of customer delay costs to assure stability of the equilibrium.[7]

Consider a stylized example with two customer types and delay cost functions:

$$C^1(t) = \frac{\alpha}{2}t^2 \quad \text{and} \quad C^2(t) = t. \tag{17}$$

Our interest is in moderate values of the parameter $\alpha \geq 0$ that model a nontrivial tension between the two customer types' delay costs. Clearly, as $\alpha \to \infty$, type 1 should get static priority, while type 2 should get static priority as $\alpha \to 0$. We assume[8] Poisson arrivals and exponentially distributed service times with type-dependent mean $m_i = 1/\mu^i$ so that $F^i(t) = 1 - \exp(-\mu^i t)$.

As a first benchmark to the Gcμ rule, consider undifferentiated service. With a single queuing class, the Gcμ rule simplifies to FIFO. The resulting M/G/1 queuing system has compound Poisson arrivals with rate $\Lambda = \lambda^1 + \lambda^2$ and a two-phase hyperexponential service distribution with mean $\mu^{-1} = \sum_{i=1}^{2}(\lambda^i/\Lambda)m_i = \rho/\Lambda$ and second moment $\sum_{i=1}^{2}(\lambda^i/\Lambda)2m_i^2$. This yields the rate γ in the Brownian approximation (16) for the workload distribution:

$$\gamma^{-1} = \mu^{-1}\frac{\rho}{1-\rho}\frac{C_a^2 + C_s^2}{2} = \frac{m_1\rho_1 + m_2\rho_2}{1-\rho}. \tag{18}$$

Under FIFO, $g(W) = W$ and Proposition 3 then yields the delay distribution of the FIFO M/G/1 system: $F^{FIFO}(t) \simeq F_W(\rho t)$, which is exact for M/M/1 if $\mu^1 = \mu^2$. The delay moments $Et_{FIFO}^k \simeq k!\rho(\rho\gamma)^{-k}$ yield[9] the total delay cost rate $DC^{FIFO} \simeq \alpha\lambda_1^1\rho^{-1}\gamma^{-2} + \lambda_1^2\gamma^{-1}$. As a second benchmark, consider differentiated service, using preemptive and nonpreemptive static priority scheduling rules, allocating types to classes. Exact expressions for the first and second delay moments are summarized in Table 1.

(i) The Gcμ Rule. Under differentiated service with two grades, the two customer types may spread their load over both service grades yielding load vectors $\lambda^1 = (\lambda_1^1, \lambda_2^1)$ and $\lambda^2 = (\lambda_1^2, \lambda_2^2)$. Under asymmetric information, classes correspond to grades and class utilization $\rho_k = m_1\lambda_k^1 + m_2\lambda_k^2$. While the index (12) is immediately applicable, it is insightful to solve (13) and (14) for the mapping g:

$$g(W) = \begin{cases} \left(\frac{u_2 W + u_3}{u_1 + u_2}, \ \frac{u_1 W - u_3}{u_1 + u_2}\right) \\ \quad \text{if } W \geq \max\left(\frac{-u_3}{u_2}, \frac{u_3}{u_1}\right), \\ (W, 0) \\ \quad \text{if } 0 \leq W < \frac{u_3}{u_1} \text{ (and thus } u_3 \geq 0), \\ (0, W) \\ \quad \text{if } 0 \leq W < \frac{-u_3}{u_2} \text{ (and thus } u_3 < 0), \end{cases} \tag{19}$$

where $u_1 = \alpha(\lambda_1^1/\rho_1^2)$, $u_2 = \alpha(\lambda_2^1/\rho_2^2)$, $u_3 = \lambda_2^2/\rho_2 - \lambda_1^2/\rho_1$. As shown in Figure 2, g is a switching curve that defines the Gcμ scheduling rule in the workload space $\mathbb{R}_+^2$: whenever the workload vector (W_1, W_2) deviates from the switching curve, serve the class that brings the workload vector back to the vector $g(W_1 + W_2)$ on the switching curve. In the case of full information, scheduling is type specific and allocates customers to classes: $\lambda_k^i = \delta_{ik}\Lambda_i$. Hence, $u_2 = 0$ and the switching curve g defines a simple *threshold rule* represented by the dashed line in Figure 2. Note that this Gcμ rule validates the practice of *expediting* jobs that have been waiting too long, thereby violating a static priority schedule and giving empirical evidence that delay costs are nonlinearly increasing. As noted earlier, static priority rules can be interpreted as two extreme cases of the Gcμ rule for $\alpha \to 0$ or ∞.

[7] The equilibrium is stable in all our examples. One should investigate mixed strategy equilibria in the unusual case that there is no Nash equilibrium in pure strategies, to which we restrict attention here.

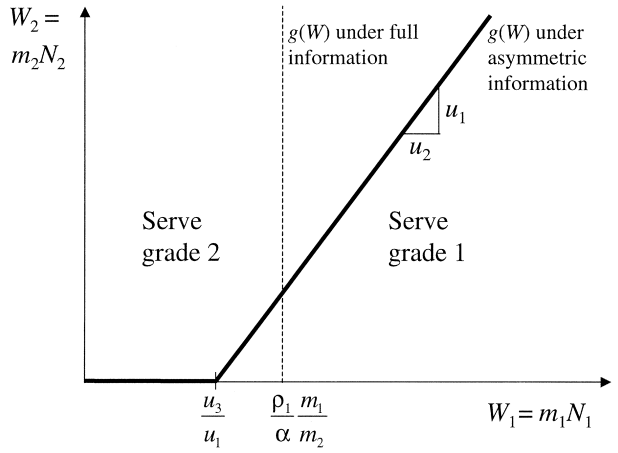
[8] Gcμ easily handles general service time distributions and more complex delay cost functions; For example, Gcμ for lead-time constraints is a generalization of longest queue policies, as shown in Van Mieghem (2000b). We assumed quadratic costs and exponential service times to allow analytic benchmarking with static priority rules. (Exact calculation of first and second moments is about as "good as it gets" for static priority rules.)

[9] Note that Et_{FIFO} is always exact, while the exact $Et_{FIFO}^2 = 2\gamma^{-2} + 2((m_1^2\rho_1 + m_2^2\rho_2)/(1-\rho))$. Our approximation $Et_{FIFO}^2 \simeq 2\rho^{-1}\gamma^{-2}$ is exact for $m_1 = m_2$ and, for general $m_1 \neq m_2$, asymptotically exact for $\rho \to 1$.

Table 1 **Moments of the Class Delays for the Gcμ Threshold Rule and for Nonpreemptive ($SPNP_1$) and Preemptive-Resume Static Priority (SPP_1) to Class 1.** $\theta = \gamma\rho_1 m_1/\alpha m_2$ **and** $\phi = m_1 m_2(\rho_1(1+\rho_1))/((1-\rho_1)^3) + m_2/m_1(2\rho_1 m_2 + m_1)(\rho_1)/(1-\rho_1))$.

	Gcμ	$SPNP_1$	SPP_1
Et_1	$\frac{\rho}{\gamma\rho_1}(1-e^{-\theta})$	$\frac{1-\rho}{\gamma(1-\rho_1)}$	$\frac{m_1\rho_1}{1-\rho_1}$
Et_2	$\frac{\rho}{\gamma\rho_2}e^{-\theta}$	$\frac{1}{\gamma(1-\rho_1)}$	$\frac{1}{\gamma(1-\rho_1)} + \frac{m_2\rho_1}{1-\rho_1}$
Et_1^2	$\frac{2\rho}{(\gamma\rho_1)^2}(1-(1+\theta)e^{-\theta})$	$\frac{2(\rho_1 m_1^2+\rho_2 m_2^2)}{1-\rho_1} + \frac{2\rho_1 m_1(1-\rho)}{\gamma(1-\rho_1)^2}$	$\frac{2\rho_1 m_1^2}{(1-\rho_1)^2}$
Et_2^2	$\frac{2\rho}{(\gamma\rho_2)^2}e^{-\theta}$	$\frac{2(\rho_1 m_1^2+\rho_2 m_2^2)}{(1-\rho_1)^2(1-\rho)} + \frac{2\rho_1 m_1}{\gamma(1-\rho_1)^3} + \frac{2}{\gamma^2(1-\rho_1)^2}$	$Et_2^{2^{SPNP_1}} + \frac{2m_2\rho_1}{\gamma(1-\rho_1)^2} + \phi$

Figure 2 **The Gcμ Rule**



Note. When customers submit to multiple service grades (boldface) it becomes a threshold rule (dashed) when customer i is optimally routed to queue class i.

(ii) The Delay Distributions and Costs. Using Proposition 3, the asymptotic delay distributions under the Gcμ threshold rule are:

$$F_1^{Gc\mu}(t|\Lambda) \simeq \Pr\left\{t_1 \simeq \frac{g_1(W)}{\rho_1} \le t\right\} = \begin{cases} F_W(\rho_1 t) & \text{if } t < \frac{1}{\alpha}\frac{\mu_2}{\mu_1}, \\ 1 & \text{if } t \ge \frac{1}{\alpha}\frac{\mu_2}{\mu_1}, \end{cases}$$

$$F_2^{Gc\mu}(t|\Lambda) \simeq \Pr\left\{t_2 \simeq \frac{g_2(W)}{\rho_2} \le t\right\} = F_W\left(\frac{\rho_1}{\alpha}\frac{\mu_2}{\mu_1} + \rho_2 t\right).$$

It is impressive how easily distributions that are intractable in exact analysis, are obtained in heavy traffic. Total delay costs $DC = \lambda_1^1(\alpha/2)Et_1^2 + \lambda_2^2 Et_2$ require the first and second delay moments, which are easily calculated and are summarized in Table 1. Denoting $f(\rho_1, \rho_2) = \mu_2(m_1\rho_1 + m_2\rho_2)\rho/(1-\rho)$, $\theta = \gamma(\rho_1/\alpha)(\mu_2/\mu_1)$ and $h(\theta) = \theta^{-1}(1-e^{-\theta})$ yields our approximation of DC^{r^*} together with bounds because $\theta \le (1-\rho)/m_2\alpha \le 1/m_2\alpha$:

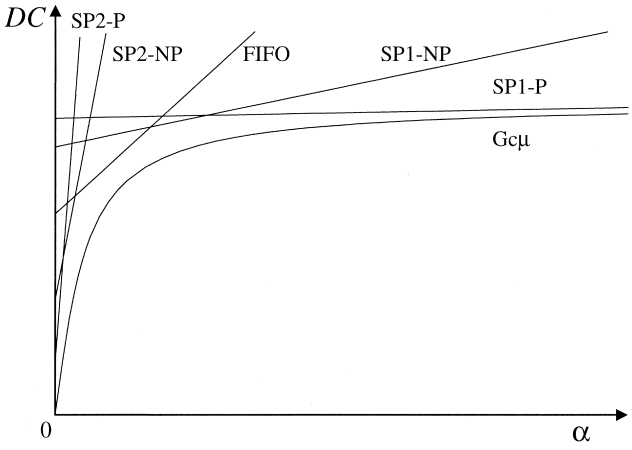
$$f(\rho_1, \rho_2)h\left(\frac{1-\rho}{m_2\alpha}\right) \le DC^{G\alpha\mu}(\Lambda) \simeq f(\rho_1, \rho_2)h(\theta) \le f(\rho_1, \rho_2). \quad (20)$$

Thus, $DC^{G\alpha\mu}$ is "sandwiched" between two convex functions and "the sandwich becomes very thin"[10] for small α, or if $\rho \to 1$, or if m_2 is large. Extremely tedious analysis shows that $DC^{G\alpha\mu}$ is indeed jointly convex in Λ for all α if $m_1 = m_2$. (Verifying convexity analytically for general $m_1 \neq m_2$ was deemed too laborious.)

While the static priority rules yield same order of magnitude delay costs, the Gcμ threshold rule balances dynamic priorities relative to α to minimize total delay cost. Obviously, delay costs are increasing in the delay sensitivity parameter α, as shown in Figure 3. All benchmark scheduling rules are linearly increasing in α, except for the Gcμ rule: the intercepts are ordered for $m_1 = m_2$, and the reverse ordering applies to the slopes. Hence, as α increases, the best benchmark rule

[10] For our examples we have $\rho \ge 0.5$ and $\alpha \le 2$ so that $h((1-\rho)/m_2\alpha) \ge h(\frac{1}{4}) = 0.8848$ for all $m_2 \le 1$, which yields a "sandwich thickness" of less than 12%.

Figure 3 **Benchmarking of Total Delay Costs *DC* As a Function of Delay Sensitivity**



switches from SPNP2 to FIFO to SPNP1 to SPP1 as intuitively expected. The interesting observation here is that it would not always be beneficial to offer differentiated services if the service provider were restricted to traditional static priority scheduling rules. The G$c\mu$ rule, however, always gives superior value (in terms of delay costs) to differentiated services.

(iii) Optimal Rates and the Value of Centralized Differentiated Service. Convexity of $DC^{G\mu}(\Lambda)$ implies that the centralized "near-optimal" vector $\Lambda^{*,\ G\mu}$ is unique for any concave value functions $(V^1,\ V^2)$. To quantify the optimal rates λ^* and total monetary value Π^* in the centralized system, we first must specify the customer gross utility functions V^i. For simplicity, assume zero operating costs ($C^S=0$) and value functions that only depend on a type's total subscription rate[11] and expected service time:

$$V^i(\lambda)=m_i\frac{\lambda^{1-\beta}}{1-\beta}, \tag{21}$$

with $0<\beta<1$. The concave rate equations (7) are easily solved numerically for the rates λ_i^*. The associated optimal profits and utilization[12] are reported in Figure 4 as a function of the delay sensitivity parameter α. The G$c\mu$ threshold rule yields both the highest monetary profit and highest utilization: It is able to serve more customers profitably.

(iv) Perfect Discriminating Prices. Numerical analysis verified that both customers' delay cost rates DC_i are locally convex in λ_i at $\lambda^{*,\ G\mu}$. Thus, Proposition 2 guarantees that the G$c\mu$ threshold rule together with $(\lambda^{*,\ G\mu},\ P^{*,\ G\mu})$ form an equilibrium[13] in the decentralized system under full information. The server captures all system value by charging type i a fixed subscription fee a_i^* and a variable price $b_i^{*,\ G\mu}=\lambda_{j\neq i}^*(\partial/\partial\lambda_i)DC_{j\neq 1}^{G\mu}$ evaluated at $\lambda^{*,\ G\mu}$, which can be calculated analytically (and shall be used for time-dependent pricing in the next section):

$$b_1^{*,G\mu}=\frac{\rho\mu_2 e^{-\theta}}{(1-\rho)}(m_1(\rho\gamma)^{-1}+m_1^2(1-(1-\rho)\alpha^{-1}\mu_2) +m_1^3\lambda_1\alpha^{-1}\mu_2+m_1^4\theta\mu_1^2) \tag{22}$$

$$b_2^{*,\ G\mu}=\frac{\alpha}{(1-\rho)\rho_1 m_1}\Bigg(-m_2^{-1}\rho\left(\frac{\rho_1}{\alpha}m_1\right)^2e^{-\theta} -(1+\rho)\frac{\rho_1}{\gamma\alpha}m_1e^{-\theta}-\gamma\rho\left(\frac{\rho_1}{\alpha}m_1\right)^2e^{-\theta} +m_2\left((1+\rho)\gamma^{-2}(1-e^{-\theta})-2\rho\frac{\rho_1}{\alpha}m_1e^{-\theta}\right) +m_2^2 2\gamma^{-1}\rho(1-e^{-\theta})\Bigg). \tag{23}$$

[11] These value functions are often used in economics because they yield demand functions with constant elasticity $1/\beta$: $V^{i'}(\lambda)=p\Leftrightarrow\lambda=(\mu_i p)^{-1/\beta}$. (When we graph results, we will fix $\beta=1/2$.) We also analyzed concave parabolic value functions. While this results in minor numeric changes it does not alter any of our conclusions.

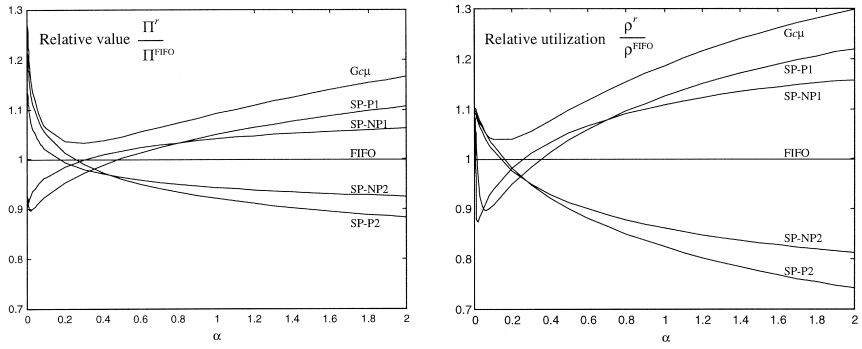
[12] Utilization decreases from $\rho\simeq 1$ for small α (type 1 is delay insensitive) to 0.4 for large α.

[13] Both customer's delay cost rates DC_i are "sandwiched" between convex functions:

$$DC_1=f(\rho_1,\ \rho_2)\left(\frac{1-(1+\theta)e^{-\theta}}{\theta}\right)\quad\text{and}\quad DC_2=f(\rho_1,\ \rho_2)e^{-\theta},$$

but DC_1 is actually not convex near $\lambda_1=0$. Yet, both π_i are unimodal convex-concave in λ_i, implying that the equilibrium λ^* is also unique.

Figure 4 **Comparing Differentiated Service Under Different Scheduling Rules to FIFO: Optimal Centralized Profits and Utilization**



6. Price and Service Discrimination Under Asymmetric Information

Section 3 showed that, under full information, the firm can choose grades, prices, and scheduling that coordinate the system so that customers self-select the centralized-optimal rates. With full information, perfect service discrimination and grade incentive compatibility is easily accomplished through type-dependent scheduling. Variable prices only had one remaining task: inducing the correct choice of aggregate rate by each type. With grade and rate IC, the server captures all system profits through perfect price discrimination.

Under asymmetric information, however, the server cannot observe the type of each job and scheduling and pricing can only be based on coarser grade information (queuing classes correspond to grades). In general, this results in imperfect price and service discrimination. Lacking sufficient information, the server can no longer perfectly align customer incentives with the optimal centralized objectives. That may lead to an equilibrium that differs from the centralized optimal (λ^*, r^*), resulting in a *coordination loss*, denoted by Δ, where $\Delta = \Pi^* - \pi^S - \sum_{i=1}^m \pi^i \geq 0$. Moreover, the service provider may no longer be able to extract all system profits and has to share them with customers. The next subsection, however, shows that, in a simplified model, scheduling and prices can perfectly coordinate the system by inducing *both* the right grade choice and the right rate choice into each grade.

6.1. Incentive Compatibility of $r^* \simeq Gc\mu$ in the "Atomistic" Model

We now describe a special case of our model, which we call the *atomistic model*, that has been introduced by Mendelson (1985) and adopted by Lederer and Li (1997) and Mendelson and Whang (1990), among others. Market segment or type i now consists of many atomistic customers who each decide individually whether or not to subscribe at an infinitesimal rate. As before, type i customers share the same delay cost function C^i but are heterogeneous in that they derive different value from processing. Similar to traditional demand curves in economics, type i customers can be ordered in decreasing value and it is convenient to endow them with a "label": the customer with label x^i receives value $v^i(x^i)\,dx^i$ from service at rate dx^i. With each customer receiving infinitesimal value, the surplus and thus fixed fee is infinitesimal ($a_i = 0$) and pure variable prices $\tilde{b}_i^*$ are socially optimal (tildes denote atomistic model). As before, customers can subscribe to multiple grades k: $dx^i = \sum_k dx_k^i$. When subscribing to grade k at rate dx_k^i, the customer incurs the self-regulating delay cost $[\mathrm{E}_\Lambda^* C^i(t_k) + \sum_{j \neq k} (dx_j^i)\,(\partial/\partial\lambda_k^i)\mathrm{E}_\Lambda^{r^*} C^i(t_j)]\,dx_k^i$ and the charge $\tilde{b}_k^* dx_k^i$, which simplify (to first order) our profit expression (11) to:

$$\tilde{b}_k^i(x^i) = v^i(x^i) - \mathrm{E}_\Lambda^{r^*} C^i(t_k) - \tilde{b}_k^*. \tag{24}$$

As before in §3, a central planner with full information makes queuing classes type specific. Let λ_i denote the label of the "marginal type i customer" who is indifferent between subscribing to grade i or not: $\tilde{b}_i^i(\lambda_i) = 0$. By

construction, $v^i(x^i)$ decreases in x^i, so that all customers with labels $x^i < \lambda_i$ have $\tilde{b}_i^i(x^i) \geq 0$ and subscribe.[14] The aggregate result of the atomistic customers' individual decisions is as before: type i rate, value, and delay cost are $\lambda_i = \int_0^{\lambda_i} dx^i$, $V^i(\lambda_i) = \int_0^{\lambda_i} v^i(x^i)\,dx^i$ and $DC_i = \lambda_i \mathrm{E}_\Lambda^{r^*} C^i(t_i)$. Thus, Proposition 1 applies verbatim to the atomistic model. Comparing, as before in Proposition 2, individual optimality ($\tilde{b}_i^i(\lambda_i^*) = 0$) with centralized optimality (7) yields the socially optimal variable price:

$$\tilde{b}_i^* = c_i(\Lambda^*) + \sum_{j=1}^m \lambda_j^* \frac{\partial}{\partial \lambda_i^i} \mathrm{E}_{\Lambda^*}^{r^*} C^j(t_j). \tag{25}$$

Hence, adopting the atomistic model moves the externality $\lambda_i^*(\partial/\partial\lambda_i)\mathrm{E}_\Lambda^{r^*} C^i(t_i)$ that the marginal type i customer inflicts on higher valued type i customers from our profit expression (11) to the price $\tilde{b}_i^*$. The results of Lederer and Li (1997) and Mendelson and Whang (1990) now directly extend to arbitrary delay cost structures and the dynamic Gcμ rule:

PROPOSITION 4. *Under asymmetric information,* $(n^* = m, \lambda^*, \tilde{b}^*, r^*)$ *is a socially optimal equilibrium in the atomistic model if marginal operating costs and service time distributions are homogeneous* (*i.e., type independent*): *The prices* $\tilde{b}^*$ *and the scheduling rule* $r^* \simeq$ Gcμ *are grade and rate incentive compatible.*

PROOF. Suppose the mechanism was not grade IC. Then, there would exist an atomistic type i customer with label $x \leq \lambda_i$, that has incentive to submit to a grade $j \neq i$ so that $\tilde{b}_j^i(x) > \tilde{b}_i^i(x)$. The term $v^i(x)$ cancels and substituting (25) into (24) yields, with $c_j(\Lambda^*) = c_i(\Lambda^*)$:

$$\mathrm{E}_\Lambda^{r^*} C^i(t_j) + \sum_{j'=1}^m \lambda_{j'}^* \frac{\partial}{\partial \lambda_j^j} \mathrm{E}_{\Lambda^*}^{r^*} C^{j'}(t_{j'})$$

$$< \mathrm{E}_\Lambda^{r^*} C^i(t_i) + \sum_{i'=1}^m \lambda_{i'}^* \frac{\partial}{\partial \lambda_i^i} \mathrm{E}_{\Lambda^*}^{r^*} C^{i'}(t_{i'}). \tag{26}$$

With equal service time distributions ($F^i = F^j$), type i jobs and type j jobs are indistinguishable from a scheduling perspective. Thus, adding an infinitesimal rate dv of either type i jobs or type j jobs to grade (=class) j impacts delay distributions, and thus expected delay costs, identically: $(\partial/\partial\lambda_j^i)\mathrm{E}_\Lambda^{r^*} C^{j'}(t_{j'})\,dv = (\partial/\partial\lambda_j^j)\mathrm{E}_\Lambda^{r^*} C^{j'}(t_{j'})\,dv$. Hence, (26) simplifies to $(\partial/\partial\lambda_j^i)DC_\Lambda^{r^*} - (\partial/\partial\lambda_i^i)DC_\Lambda^{r^*} < 0$, which means that total delay costs $DC_\Lambda^{r^*}$ would decrease by reallocating an infinitesimal rate dv of type i from queue class i to queue class $j \neq i$. This would contradict the optimality of r^* for any flow vector Λ. Hence, there exist no x, i, or $j \neq i$ that satisfies $\tilde{b}_j^i(x) > \tilde{b}_i^i(x)$. Thus, the mechanism is grade IC and, by construction of the prices $\tilde{b}^*$ so that $\tilde{b}_i^i(\lambda_i^*) = 0$, also rate IC. □

With "one-dimensional types" that are differentiated along a single dimension (say value), Clarke (1971) and Groves and Loeb (1975) showed that charging a customer his or her "full externality," which is the difference between Π^* and the optimal system profits that obtain when that customer would not be present, is incentive compatible. With homogeneous service times, types differ in value and delay costs, but only delay costs create externalities so that types are essentially one-dimensional. In addition, with atomistic customers (25) equals the full externality and hence yields incentive compatibility, in agreement with Clarke-Groves-Loeb. The next section discusses the case of heterogeneous service times,[15] monopoly pricing and nonatomistic customers.

6.2. Incentive-Compatibility in the "General" Model

In general, the optimal design problem under asymmetric information with multidimensional types is intractable and involves a calculus of variation problem over all possible pricing functions and scheduling rules. Our mode of approximate analysis, however, makes the scheduling problem tractable: in §4 we specified the Gcμ rule for the general case of asymmetric

[14] Hence, in the atomistic model, variable prices are no longer perfectly discriminating because all type i customers with labels $x_i < \lambda_i$ have a higher "reservation price" than $\tilde{b}_i^*$ and enjoy a customer surplus.

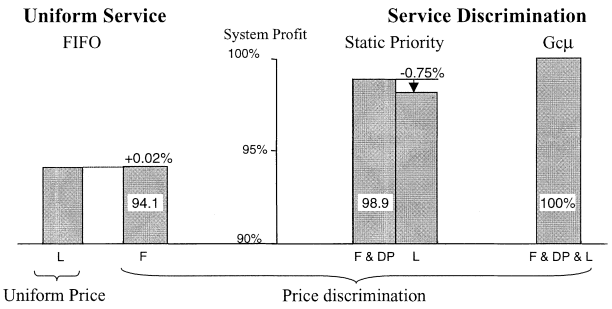
[15] Lederer and Li (1997) argue that there is no information problem if a job's type can be ascertained by its processing requirements. That is the case for deterministic service times, or if the service times distributions of different types do not overlap. If they do, one must resort to service-time-dependent pricing as in §6.2.

information. In practice, the calculus of variation pricing problem is approximated by an optimization over a parametrized family of pricing functions. We consider two pricing families: (1) Single-plan pricing charges each customer a two-part tariff $P^{SP}(\lambda) = a + \sum_{k=1}^{n} b_k \lambda_k$ for its subscription vector λ. (Without the fixed subscription fee, a, this reduces to linear pricing.) One could consider multipart tariffs, but typically most gains are captured by a two-part tariff relative to linear pricing, as will be illustrated shortly. (2) Under multiplan pricing, each customer must choose one plan from a menu that specifies p two-part tariffs and is charged $P^{MP}(\lambda) \in \{a_l + \sum_{k=1}^{n} b_{lk}\lambda_k : l = 1, \ldots, p\}$.

The executable proposal under asymmetric information becomes: (i) Fix a number of service grades n and characterize the G$c\mu$ rule under asymmetric information for the delay cost functions at hand. (ii) Calculate the associated delay distributions and total expected delay cost DC^*. (iii) Fix a pricing plan and vectors a and b, and find the associated Nash equilibrium rates $\lambda^i(a, b)$ by solving the system of first-order constraints (10). Calculate corresponding customer and server profits $\pi^i(a, b)$ and $\pi^S(a, b)$ via (1) and (5). (iv) Optimize over vectors a and b by iterating step (iii) to maximize either system profits $\Pi = \pi^S + \sum_i \pi^i$ for social pricing or server profits π^S for monopoly pricing. (v) Optimize over number of grades n.

To investigate the possibility of perfect coordination it is reasonable to first restrict the search to perfect service-discriminating mechanisms by setting $n = m$ and imposing the grade IC conditions (11). This simplifies the optimization because now $\lambda^i_{k \neq i} = 0$ leaving only m unknown rates, but at the expense of adding the grade IC constraints. One would use the perfectly discriminating prices b^* from Proposition 2 as the initial conditions for the variable prices b. In the special case that the centralized-optimal rates λ^* are still an equilibrium (as in the atomistic model), perfect coordination is achieved and no further optimization is needed under social welfare maximization: $n^* = m$, λ^*, and b^* are socially optimal. If, however, another rate vector than λ^* provides the optimal grade IC equilibrium, then one should also optimize over the class of not-grade-IC mechanisms because the best imperfect discriminating mechanism

Figure 5 **Relative System Profits Under Social Welfare Maximization Under Full Information with Perfect Price Discrimination (F) and Under Asymmetric Information with Dual Plan (DP) and Linear Pricing (L) for Our Example with $\alpha = 0.1$ and $\mu_i = 1$**

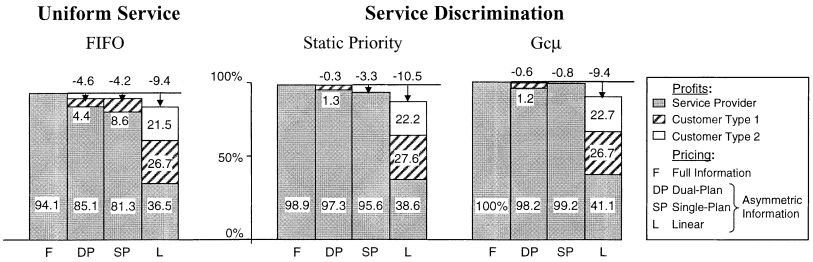


may conceivably outperform the best perfect service discriminating mechanism. Even if (λ^*, b^*) form a socially optimal equilibrium, the server may extract a larger profit π^S with another equilibrium under monopoly pricing. In the remainder of this section, we will investigate whether a perfect-service discriminating equilibrium is actually optimal or not.

Until to date, we have not been able to present a simple price expression as in Proposition 4 for the not-atomistic model under asymmetric information. (Clearly, charging the full externality does the trick, but that expression is complex.) The reason is the presence of the term $\lambda_i^*(\partial/\partial\lambda_i)\mathrm{E}^{r^*}_{\Lambda} C^i(t_i)$ in (11) in the general model, rather than in the price term (25) in the atomistic model. In addition, with finite-sized customers, the "full externality" typically differs from its differential approximation that appears in the central-optimal prices b^* and it seems unlikely that the prices b^* could be shown in general to be rate IC (although our results below suggest they may be). Therefore, we will restrict attention in the remainder to our two-type example of §5, but now under asymmetric information, an highlight some insights.

• In our example, *the prices b^* remain grade and rate IC with the Gcμ rule in the general model under social welfare maximization, but not under monopoly maximization.* The optimized system profits for our example from §5 under social welfare maximization are shown in Figure 5 assuming homogeneous service times $\mu_1 = \mu_2 = 1$

Figure 6 **Relative Agent Profits Under Server Profit Maximization for Four Different Pricing Strategies in our Example with $\alpha = 0.1$ and $\mu_i = 1$**



and $\alpha = 0.1$. As expected, using suboptimal traditional scheduling rules yields a positive coordination loss, which is exacerbated under asymmetric information. With the dynamic G$c\mu$ rule, however, no value is lost and the prices $b_i^{*,\,Gc\mu}$ of (22) and (23) perfectly coordinate the system (adopting the G$c\mu$ rule as a proxy for the unknown optimal rule r^* to verify numerically that $\Delta \simeq \Delta|_{r^* \simeq Gc\mu} = 0$ for all $\alpha \in [10^{-4}, 10^4]$). This suggests that Proposition 4 may well extend to not-atomistic customers. Under server profit maximization (monopoly pricing), however, the G$c\mu$ rule incurs a coordination loss as shown in Figure 6.

• The example shows that *the addition of a fixed subscription fee has the largest gain relative to linear monopoly pricing*. A single two-part tariff allows the service provider to extract all system profits under differentiated service, more than doubling his profit compared to linear pricing (Figure 6). Under social welfare maximization, however, single-plan two-part tariffs in essence reduce to linear pricing because the specific allocation of profits is immaterial.

• *Negative feedback in the* G$c\mu$ *induces grade IC*. It is not surprising that static preemptive priority rules suffer from an increased coordination loss under asymmetric information. If they were to use the perfect discriminatory prices b^*, the lowest grade corresponding to the lowest priority service would be free because it imposes no externality cost on higher grades. Costless service may tempt the customers who in the centralized system should be using higher priority grades at higher price to "cheat" and also submit to the lowest priority grade. The G$c\mu$, rule, however, uses negative feedback control to *actively discourage cheating*: it charges positive prices and "threatens" to adjust its scheduling to counteract cheating. Indeed, assume customer 2 were to shift some of its load to grade 1. The G$c\mu$ rule would increase its threshold (Figure 2), which would increase the QoS of grade 2 and reduce grade 1's QoS to discourage customer 2 from cheating. This shows how dynamic scheduling can improve price discrimination.

• *Service-time dependent pricing is needed to induce grade IC and perfect coordination with heterogeneous service times*. It is surprising that the perfect discriminatory prices b^*, deduced under full information, remain optimal with the G$c\mu$ rule under asymmetric information in Proposition 4 and in our example. It is not true that those simple prices are always optimal: Under heterogeneous service times and social welfare maximization, the G$c\mu$ rule incurs a coordination loss if $\alpha < 0.6$ ($\forall \mu_i \in [0.1, 10]$). For example, for $\alpha = 0.1$ and $\mu = (10, 2)$, the prices $b^{*,\,Gc\mu} = (0.015, 0.376)$ are no longer grade IC.

The perfect discriminatory price $b_i^{*,\,Gc\mu}$ specifies the charge for a type i job, which is known to take an average of $m^i = 1/\mu^i$ time-units of the server. Typically, shorter jobs will be priced lower so that under asymmetric information a type with long jobs ($m = 1/2$ in the example with $b_2^{*,\,Gc\mu} = 0.376$) may find it better to use grades designed and priced for types with shorter jobs ($m = 1/10$ with price 0.015). To counteract this, Mendelson and Whang (1990) proposed a pricing scheme that is quadratic in the actual service time for static-priority rules and linear delay costs. With service *time dependent* (TD) pricing, customer i anticipates a variable price $Eb_k(\tau^i)$ for grade k, giving the service provider additional control to counteract cheating. For TD prices to be perfectly coordinating, they must be grade IC and

coincide with the perfect discriminatory prices in expectation to induce rate IC: $\mathrm{E}b_i^{r^*}(\tau^i) = b_i^*$.

In our example, inspired by (22) and (23), which can be written as $b_i^{*Gc\mu} = \sum_{j=0}^{4} \beta_j^i m_i^j$, we propose the TD prices: $b_i^{Gc\mu}(t) = \max\{b_i^{*Gc\mu}, \sum_{j=0}^{4} \beta_j^i (t_i^j/j!)\}$ for which $\mathrm{E}b_i^{Gc\mu}(\tau^i) = b_i^{*Gc\mu}$, so that these prices are perfectly coordinating if they are grade IC. For our example with $\alpha = 0.1$ and $\mu = (10, 2)$ types 2's price if it were to choose grade 1 now becomes $\mathrm{E}b_1^{Gc\mu}(\tau^2) = 5.217$, which is much larger than type 1's price for that grade, $b_1^{*,\,Gc\mu} = 0.015$, thereby discouraging type 2 from cheating. We verified that these TD prices are indeed grade and rate IC and thus perfectly coordinating for all μ_i and α in $[10^{-1}, 10]$.

• The example shows how *service-time dependent prices can be designed from expanding the perfect discriminatory prices in terms of the service times*. Notice that many expansions may work[16] so that coordinating TD prices need not be unique. On the other hand, we have no guarantee that coordinating TD prices always exist.

• *Multiplan pricing may further improve system performance*. It allows more control variables and will not reduce system profits because it includes single-class pricing as a special case. Under social welfare maximization, dual plan pricing can improve system performance by giving customers better incentive to truthfully reveal their type. For static priority rules, this can eliminate the increased coordination loss due to asymmetric information (Figure 5).

• *Restricting attention to perfect service discrimination* (*grade IC*) *with multiplan pricing may decrease server profits relative to single-plan pricing*. For example, under server profit maximization (Figure 6), dual plan pricing may increase server profits under FIFO and static priority rules, while under Gcμ making the server worse of than single-plan pricing. (For other α parameters, the reverse can happen.) It may be surprising that multiplan pricing may benefit customers at the expense of the service provider. The culprit is the restriction to perfect service discrimination and offering a two-part tariff $P_k(\lambda) = a_k + b_k\lambda_k$ associated with each service grade k. While multiplan pricing adds more control variables to the server, it also imposes discrete choice on the customers, which adds more grade incentive compatibility constraints. Indeed, multiplan pricing is more intricate than single-plan pricing for which the equilibrium vectors λ^i solve the first-order differential conditions (10). It involves the calculation of equilibrium vectors λ^i and corresponding customer profits π^i in each subgame that assumes a particular customer choice vector of plans, and the identification of a (subgame perfect) Nash equilibrium choice vector, which adds incentive compatibility constraints in the form of discrete choice conditions. The result of more variables and more constraints is case specific: server profit may increase or decrease relative to single-plan pricing (as it does in our example depending on the value of parameter α).

For our two-type example, each subgame requires calculating customer equilibrium rates $\lambda^i_{k_1k_2}$ and profits $\pi^i_{k_1k_2}$ by solving the first-order conditions (10) assuming customer i chooses plan k_i. Notice that the scheduling rule reduces to FIFO if both types choose the same plan. Denoting plan 0 for not subscribing, profits in all subgames can be summarized in a pay-off matrix:

$$\begin{bmatrix} (0, 0) & (0, \pi^2_{01}) & (0, \pi^2_{02}) \\ (\pi^1_{10}, 0) & (\pi^1_{11}, \pi^2_{11}) & (\pi^1_{12}, \pi^2_{12}) \\ (\pi^1_{20}, 0) & (\pi^1_{21}, \pi^2_{21}) & (\pi^1_{22}, \pi^2_{22}) \end{bmatrix}.$$

The equilibrium customer plan and rate choice for the multiplan price menu must form a Nash equilibrium choice vector $(k_1^*, k_2^*) = (1, 2)$, which adds the discrete choice IC constraints that $\pi^1_{12} \geq \max(0, \pi^1_{22})$ and $\pi^2_{12} \geq \max(0, \pi^2_{11})$. The profit-maximizing dual pricing plan menu, then, is found by optimizing $\pi^S = P_1(\lambda^1_{12}) + P_2(\lambda^2_{12})$ over $\{a_1, b_1, a_2, b_2\}$ and defines a subgame perfect Nash equilibrium to our decision problem.

7. Concluding Remarks

This article presents theory and tools to study social and monopoly pricing of heterogeneous customers, each wanting a specific service and each having a delay sensitivity for that service. Our main results can be

[16] In our example, the terms β_j^i contain θ and ρ, which can actually be further expanded into m_i, yielding another, but more complex, TD price. Also, the term in m_2^{-1} is captured into β_0^2 because no moment $\mathrm{E}(\tau^2)^k$ yields m_2^{-1}.

summarized as:

1. From a modeling perspective, we introduce delay cost *curves* that allow a flexible description of quality sensitivity.

2. We propose a comprehensive executable approach that analytically specifies scheduling, delay distributions and prices for arbitrary delay sensitivity curves. The tractability of this approach derives from porting heavy-traffic Brownian results into the economic analysis. The G$c\mu$ scheduling rule that emerges is dynamic so that in general service grades need not correspond to a static priority ranking.

3. We introduce the notions of *grade* and *rate* incentive compatibility to study this system under asymmetric information and establish them for G$c\mu$ scheduling when service times are homogeneous and customers atomistic. We illustrate with a benchmarking example and extend to time-dependent and multiplan pricing to strive for incentive compatibility with heterogeneous service times and not-atomistic customers.[17]

[17] My brother Piet Van Mieghem introduced me to differentiated QoS in communication networks and motivated this study. I am also grateful to Philipp Afèche, Mike Harrison, and Rakesh Vohra for many stimulating discussions, and to area editor Paul Glasserman and two anonymous referees for helpful questions and suggestions.

Appendix. Intuition Behind the G$c\mu$ Rule and Proposition 3

The G$c\mu$ rule is founded on three facts from queuing theory that are best described in terms of the total workload process W, which measures the total amount of work (in time units) that is waiting: $\mathrm{E}[W|N] = \sum_{j=1}^{q} N_j/\mu_j$. First, total workload is invariant for scheduling rules that do not idle when there is work present in the system. Indeed, new arrivals contribute to an increase in $\mathrm{E}W$ at (average) rate ρ, while serving drains $\mathrm{E}W$ at rate 1 regardless of which class is served, yielding a net decrease of $\mathrm{E}W$ at rate $1-\rho$, as long as there is work present. The scheduling rule, however, does impact how this total workload W is distributed over the different classes: when serving class k, its average *class* workload $\mathrm{E}W_k$ decreases at rate $1-\rho_k$, while arrivals increase the average workload of any other class j at rate ρ_j. The two other facts follow in heavy traffic: class workloads "live on a faster timescale" than the total workload process. Indeed, as $\rho \to 1$, total workload hardly changes, while class workloads keep changing at a finite rate. At the timescale of the total workload, it is as if one can almost instantaneously shift workload away from one class to the other classes by serving that class for an infinitesimal amount of time while the total workload is unchanged. Third, in well-behaved heavy traffic limit systems, the class workload process "converges": W_k/W approaches a W-dependent constant. In effect, scheduling switches among classes precisely so as to keep $(W_1, W_2, \ldots, W_q)$ close to the point $g_k(W)$ on the switching curve. Aside from fast, but small disturbances around the switching curve, the W_k thus follow the W movement and change very slowly. Meanwhile many class k jobs flow through the system at rate λ_k while $(W_1, W_2, \ldots, W_q)$ remains relatively constant. "It is as if a job takes a *snapshot* of the network when it enters and all queues remain at that same value during the job's sojourn throughout the network" (Reiman 1982, p. 413). In Van Mieghem (1995), we apply Little's law to yield the state-dependent delay:

$$t_k \overset{\text{distribution}}{\simeq} \frac{N_k}{\lambda_k} \overset{\text{distribution}}{\simeq} \frac{W_k}{\rho_k}. \tag{27}$$

Using (27), the instantaneous delay cost rate can be expressed in terms of class workloads as $\sum_{i,\ k} \lambda_k^i C^i(w_k/\rho_k)$. The intuition behind the G$c\mu$ rule then is to "distribute" the total workload $W = \sum_{k=1}^{n} W_k$ over the different classes such that this delay cost rate is minimized at each point in time. This greedy cost-minimizing allocation of total workload W to class workloads is found by solving (13) and (14) and represented as $W_k = g_k(W)$. As a positive sum of convex functions, $\sum_{i,k} \lambda_k^i C^i(w_k/\rho_k)$ is convex so that W_k solves the sufficient first-order conditions:

$$\sum_{i=1}^{m} \frac{\lambda_k^i}{\rho_k} c^i\left(\frac{W_k}{\rho_k}\right) - v_k = \sum_{i=1}^{m} \frac{\lambda_{k'}^i}{\rho_{k'}} c^i\left(\frac{W_{k'}}{\rho_{k'}}\right) - v_{k'}$$
$$\forall \text{ classes } k \text{ and } k', \tag{28}$$

with $W_k v_k = 0$ and $\sum_k W_k = W$, where $v_k \geq 0$ is the Lagrange multiplier on the nonnegativity constraint of w_k. Recognizing that $W_k/\rho_k = N_k/\Lambda_k$, shows that the G$c\mu$ rule with index[18] (12) attempts to implement the first-order conditions (28). Finally, using (27), we have that the delay distribution $F_k^{Gc\mu}(t) = \Pr(t_k \leq t) \simeq \Pr(g_k(W) \leq \rho_k t) = F_W(g_k^{-1}(\rho_k t))$. Notice that this approach is consistent with the conservation law: overall average delays $\sum_k (\rho_k/\rho) t_k$ equal the FIFO delay $t^{FIFO} \simeq W/\rho$. □

[18] Given (27), an alternative implementation of the G$c\mu$ rule keeps track of arrival times and replaces N_k/Λ_k in the index (12) by the "age" A_k of the oldest job in each class.

References

Afèche, P., H. Mendelson. 2000. Market structure in congestible tandem networks. Technical report, North-western University, Evanston, IL.

Ayhan, H., T. L. Olsen. 2000. Scheduling of multi-class single-server queues under nontraditional performance measures. *Oper. Res.* **48**(3) 482–489.

Bradford, R. M. 1996. Pricing, routing, and incentive-compatibility in multiserver queues. *Eur. J. Oper. Res.* **89** 226–236.

Cachon, G. P., P. T. Harker. 1999. Service competition, outsourcing and co-production in a queuing game. Technical report, working paper, Duke University, Durham, NC.

Clarke, E. 1971. Multipart pricing of public goods. *Public Choice* **11** 19–33.

Courcoubetis, C. 1998. Pricing and economics of networks. Tutorial presented at IEEE InfoCom Conference, San Francisco, CA. ⟨www.ics.forth.gr/~courcou⟩.

De Vany, A. S., T. R. Saving. 1983. The economics of quality. *J. Political Econom*. **91**(6) 979–1000.

Dolan, R. J. 1978. Incentive mechanisms for priority queuing problems. *Bell J. Econom.* **9**(2) 421–436.

Gibbens, R. J., F. P. Kelly. 1999. Resource pricing and the evolution of congestion control. *Automatica* **35** 1969–1985.

Groves, T., M. Loeb. 1975. Incentives and public inputs. *J. Public Econom.* **4** 211–226.

Ha, A. Y. 1998. Incentive-compatible pricing for a service facility with joint production and congestion externalities. *Management Sci.* **44**(12) 1623–1636.

——. 1999. Optimal incentive-compatible pricing for processor sharing queues with customer-chosen service requirements. Technical report, Yale School of Management, New Haven, CT.

Knutsen, N. C. 1972. Individual and social optimization in a multi-server queue with a general cost-benefit structure. *Econometrica* **40** 515–528.

Lederer, P. J., L. Li. 1997. Pricing, production, scheduling and delivery-time competition. *Oper. Res.* **45**(3) 407–420.

Lippman, S. A., S. Stidham. 1977. Individual versus social optimization in exponential congestion systems. *Oper. Res.* **25** 233–247.

Loch, C. H. 1991. Pricing in Markets Sensitive to Delay. Ph.D. thesis, Stanford University, Stanford, CA.

Lui, F. 1985. An equilibrium queuing model of bribery. *J. Political Econom.* **93** 760–781.

Maglaras, C., J. A. Van Mieghem. 2000. Admission and sequencing control under delay constraints with applications to GPS and GLQ. Technical report, Northwestern University, Evanston, IL.

Masuda, Y., S. Whang. 1999. Dynamic pricing for network service: Equilibrium and stability. *Management Sci.* **45** 857–869.

Mendelson, H. 1985. Pricing computer services: Queueing effects. *Comm. ACM* **28**(3) 312–321.

——, S. Whang. 1990. Optimal incentive-compatible priority pricing for the M/M/1 queue. *Oper. Res.* **38** 870–883.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37** 15–24.

Rao, S., E. R. Petersen. 1998. Optimal pricing of priority services. *Oper. Res.* **46**(1) 46–56.

Reiman, M. I. 1982. The heavy traffic diffusion approximation for sojourn times in jackson networks. R. Disney, T. Ott, eds., *Applied Probability-Computer Science*, *The Interface* (*Volume II*). Birkhauser, Boston, MA, 409–422.

Reitman, D. 1991. Endogeneous quality differentiation in congested markets. *J. Indust. Econom.* **39**(6) 621–647.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized *cμ* rule. *Ann. Appl. Prob.* **5**(3) 809–833.

——. 2000a. Delay distributions under dynamic scheduling rules. Working paper, Northwestern University, Evanston, IL.

——. 2000b. Scheduling with delay constraints: A short proof of the optimality of generalized longest queue (GLQ). Technical report, Northwestern University, Evanston, IL.

Accepted by Paul Glasserman; received on April 8, 1999. *This paper was with the author* 3 *months and* 3 *weeks for* 2 *revisions.*